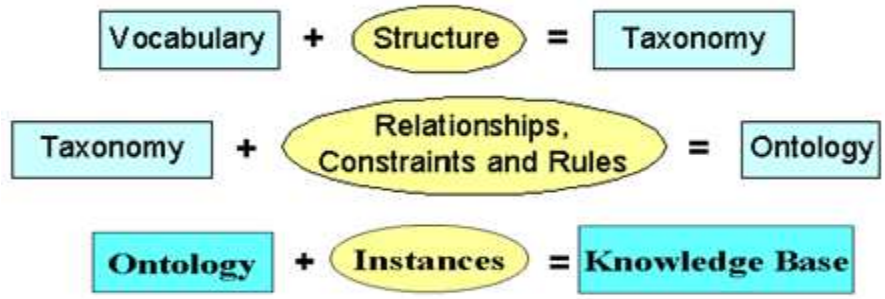
Socio-Ontological Structure of Health Domain in Tamil: A knowledge-based system

Rajendran Sankaravelayuthan & S. Veera Alagiri
Amrita Vishwa Vidyapeetham, Coimbatore

Siddha medicine is one of the traditional and effective systems of medicine that originated in Tamil Nadu. This system is considered one of the oldest and most prevalent medicine systems in India. The term Siddha refers to masters with excellence in physical and spiritual practices, and they are the early practitioners of Siddha medicine. Most of the Siddha literature is written in Tamil, and it requires proficiency in the Tamil language for the right understanding and effective practice. The best practices of the Siddha system are most needed for the well-being of humankind, specifically in the modern lifestyle. This requires the right understanding and better representation of the knowledge conveyed through the Siddha system, which can be easily accessed by the end user, web applications, and automated intelligent systems. One of the most appropriate ways of representing this traditional knowledge along with aforesaid requirements is using ontologies. Ontology adapts formal representation methods to describe the facts through which one can share the domain information without any ambiguities that can be interpretable by the machines. This project proposes a Siddha domain ontology constructed in the native Tamil language. The interpretations of this ontology are also described in English in order to access and interpret the knowledge by any application.

Simple ontological structure will keep the society away from the task. So, it is proposed here to make use of the knowledge shared by the society in building the ontological structure of health domain. The knowledge shared by the society will be taken into account while building up the proposed knowledge-based system, Socio Ontological Structure proves to be extremely useful in the representation of Health domain. (The chief investigator himself visualized it while treating his patients.) The socio-linguistic structure coupled with natural language processing (NLP) can contribute to this. The ontological representation of health domain in Tamil incorporating the outlook and interest of the society is the need of the day. English has developed general ontological structure as well other specific ontological structure including health domain. But developing socio-ontological structure is a new venture.

Ontology is based on the principle that the lexical items in a semantic domain have shared features which help us to link them by means of lexical semantic features.


Vocabulary + Structure = Taxonomy
Taxonomy + Relationships, Constraints and Rules = Ontology
Ontology + Instances = Knowledge Base

For instance, given the word *cuNTeli* '*mouse',* a proper although minimal representation of its meaning requires distinguishing the sense of 'small rodent' from the one of 'small pointing device for computers'. Moreover, the same representation should be able to capture the fact that being a rodent entail being a mammal, as well as the fact that the sense of *mouse* as 'small rodent' shares with the meaning of other words such as *naay* 'dog', or *puunai 'cat)'* the fact of being subtypes of mammal. Aristotelian thought of classify things in the world based on the principles of genera and species is proved to be a highly useful knowledge representation. Ontology is an explicit specification of a conceptualization. As a term borrowed from philosophy, ontology means a systematic account of 'existence'. Artificial Intelligent systems (AI) deliberates that what "exists" can be represented in a program code. They try to represent knowledge of a domain in a declarative formalism; the set of objects that can be represented is called the universe of discourse.

2. Present Status of Ontological structure in Tamil

Rajendran has developed an onto-thesaurus for Tamil vocabulary (1982, 2001), which is founded on the theory of componential analysis of meaning propounded by Nida (1975), Indian tradition of *nighaNTu* and Aristotelian principle of genera and species is enhanced to suit the present purpose. Rajendran (2001) has classified the concepts represented by the vocabulary into four types: entities, abstracts, abstracts and relationals following Nida (1975). The entities consist of referential meanings of concrete concepts, events consist mainly of verbs and verbal nouns and abstracts consist mainly of adjectives and adverbs apart from abstract nouns and relational consists of functional words including postpositions, connectives and coordinators.

3. Two types of issues

Ontology by default implies a thesaurus which in its wider sense is a classification of words by concepts, topics, or subjects. The present socio-ontology is an elaborate structure of health domain. There are two kinds of issues:

- Linguistic issues
- Computational issues

Linguistic Issues

It involves mainly the following four tasks:

1. Developing a data base for Health domain in Tamil

2. Establishing semantic domains and sub domains of health domain based on semantic principles and linking the lexical items with various semantic and lexical relations.

4. Apart from linking health oriented vocabulary by basic semantic or lexical relations such as synonymy, hyponymy-hyperonymy, meronymy-holonymy, compatibility, and incompatibility, specific relations establishing knowledge representation of health domain will be attempted

Computational Issues

It involves mainly the following three tasks:

1. Conversion of linguistic data base into computer accessible format.
2. Preparation of a tool to provide the facilities for augmenting, entering and editing the raw data, and classifying the lexical items in a semi-automatic way.
3. Creation of user friendly interfaces for accessing the knowledge based socio-ontological system in a simple manner.

4. Structuring of health Vocabulary by Lexical Relations

There are at least four lexical or meaning relations by which lexical items can be linked or related to one another in the ontological structure of Tamil vocabulary.

Synonymy: e.g. maruttuvar 'doctor': vaittiyar ‘doctor’

Hyponymy-hypernymy: e.g. maruttuvam ‘medicine*:* cittamaruttuvam ‘siddhaa medicine’

Meronymy-holonymy: e.g. *uTal* ‘body’: *kAl ‘hand and kaal* ‘leg’;

Compatibility: e.g. kaNnooy ‘eye disease: kaNcivappu ‘redness of eye’

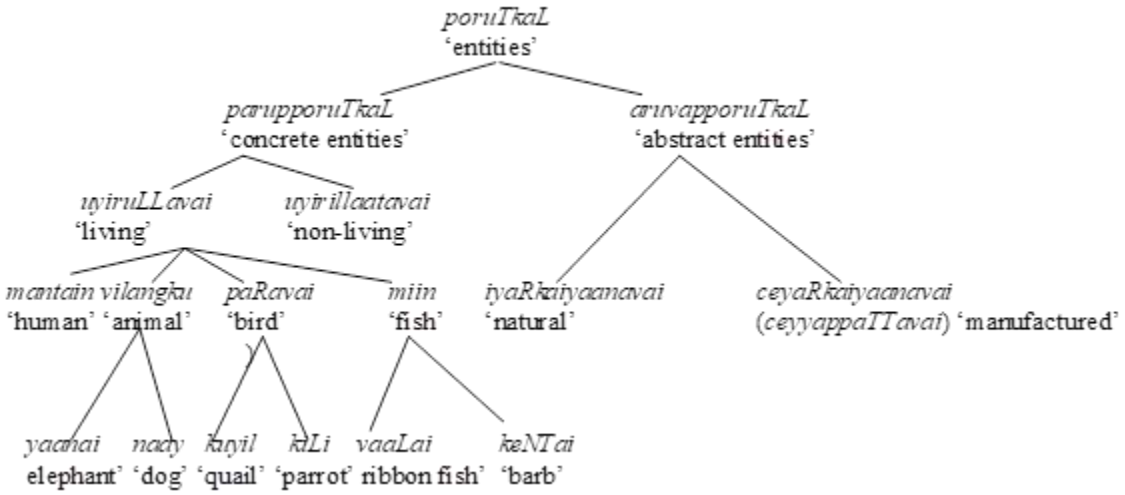
Incompatibility: *cittamaruttuvam*: *aayurveedic*; incompatibility leads to the relation called opposition which culminates into many types which are discussed below.

Lexical Inheritance

Hypernymy-hyponymy and meronymy-holonymy assure a lexical item to inherit semantic features.

Hierarchies

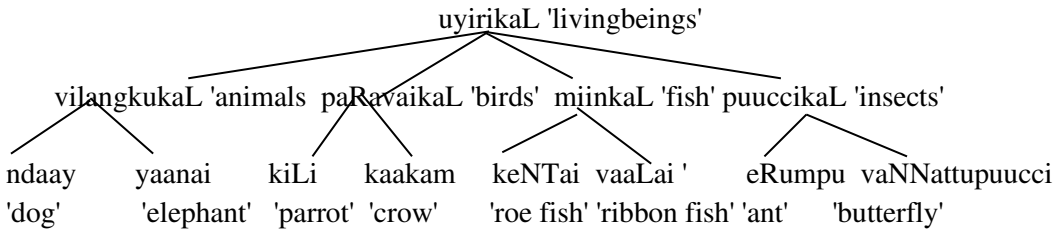
Hierarchies (Cruse 1986:112-118) are of two types: branching hierarchies and non-branching hierarchies.



Taxonomic Hierarchies

The hierarchies are the outcome of the hypronymy-hyponymy relation between lexical items. The sequence of hyperonymy-hyponymy relation leads the hierarchical structure of a set of vocabulary items which shows this pair of relations among themselves. Taxonomic hierarchies are more liberal than hypronymy-hyponymy hierarchy. The following is an example:

:

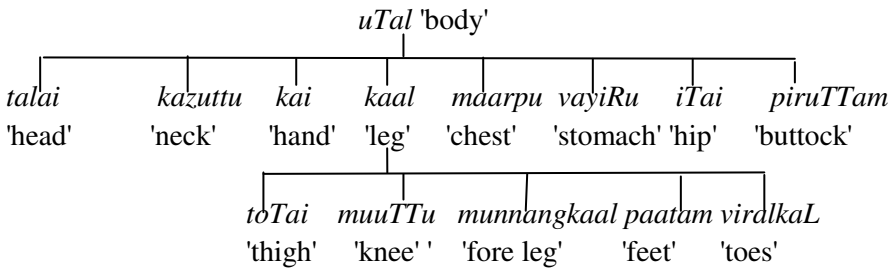


The health vocabulary will be formatted into taxonomic hierarchies which give a hierarchical stature to health vocabulary.

Meronymic Hierarchies

In addition to taxonomic hierarchies meronymic hierarchies also will be formulated which will complement the taxonomic hierarchies.

Meronymic hierarchies are the result of meronymy-holonymy relation shown by the lexical items. The meronymy-holonymy relation also gives hierarchical structure to a set of vocabulary. The following is an example.



Non-branching Hierarchies

Non-binary opposition leads to a number of types of non-branching hierarchies. They are listed below:

Bipoles: The bipoles is the simplest kind of linear structure found in a pair of opposites. They are simply oppositions which we have discussed earlier (Cruse, 2000: 189). e.g. niiLamaana ‘long’: kuTTaiyaana ‘short’, viraivaaka ‘fast’: metuvaaka ‘slowly’.

Bipolar chains: The bipolar chains imply a scale on which a pair of opposites operates denoting different degrees of the property. They have implicit superlative terms of opposite polarity at each end of the scale (Cruse, 2000: 189). The following is the example: mikanuNNiya 'very minute', nuNNiya 'minute', ciRiya 'small', periya 'big' mikapperiya 'very big'

Monopolar Chains: In monopolar, there is no sense that the terms at the ends of the chain are oriented in opposite directions. Degree, stages, measures, ranks and sequences come under monopolar chains (Cruse, 2000: 190).

Degrees: Degrees incorporate as part of their meaning different degrees of some continuously scaled property such as size or intensity (Cruse, 2000: 190). e.g. paaRai 'mound': kunRu 'hillock': malai 'hill' : maamalai 'mountain'

Stages: Stages are points in a life cycle of something and normally involve the notion of progression (Cruse 2000: 190). e.g. kuzandtaip pruvam 'child stage': vaalipap paruvam 'young stage' : mutir paruvam 'adult stage': mutu paruvam 'old stage'.

Measures: Measures are based on part-whole relationship, with each whole divided into a number of identical parts (Cruse, 2000: 190). cekaNTu 'second', miniT 'minute', maNi 'hour', naaL 'day', vaaram 'week' maatam 'month', aaNTu 'year'.

Ranks: The lexical items under ranks entails a sequential order which is not gradual (Cruse, 2000. 191). e.g. virivuraiyaaLar 'lecture', mutunilai virivuraiyaaLar 'senior lecture', iNaippeeraaciriyar 'reader', peeraaciriyar 'professor'. Lyons (1977: 290) includes numeral under rank as an unique type. E.g. onRu ‘one’, iraNTu ‘two’ .. pattu ‘ten’, irupatu ‘twenty, … nuuRu, irunuuRu, …

Sequences: Sequences are lexical items which are ordered but do not bear increasing property as in the case of previous chains (Cruse, 2000: 191). e.g. iRandta kaalam ‘past tense’, nikaz kaalam ‘present tense’, etir kaalam ‘future tense’; kaar kaalam 'rainy season' kuuLir kaalam 'cold season', munpani kaalam 'early cold season', pinpanik kaalam 'late cold season', iLaveeniR kaalam 'milder hot season', mutuveeniR kaalam 'hot season'.

Cyclical Sets or Cycles: The sequential lexical item (Cruse, 1986:187-190) can entail a cyclical order of time in the natural arena. e.g. kaarkaalam 'rainy season' kuutirkaalam 'cold season', munpanikkaalam 'early cold season', pinpanikkaalam 'late cold season', iLaveenirkaalam 'milder hot season', mutuveenirkaalam 'hot season'; canavari 'January', pepravari 'February', maarc 'March', eeppiral 'April', mee 'May', juun 'June'…}; njaayiRu 'Sunday', tingkaL 'Monday', cevvaay 'Tuesday' , putan 'Wednesday', viyaazan 'Thursday', veLLi 'Friday', cani 'Saturday'.

Propositional series or grids

The units of a gird is the cell, which consists of four lexical items, any one of which must be uniquely predictable form the remaining three. The followings are the examples of cells: pacu 'cow': kanRu 'calf' :: kutirai 'horse' : kuTTi 'foe' ; paampu 'snake' : kunjcu 'young one' :: palli 'lizard' : kunjcu 'young one' ; yaanai : piLiRu 'trumpet' :: fox: uuLaiyiTu 'howl'.

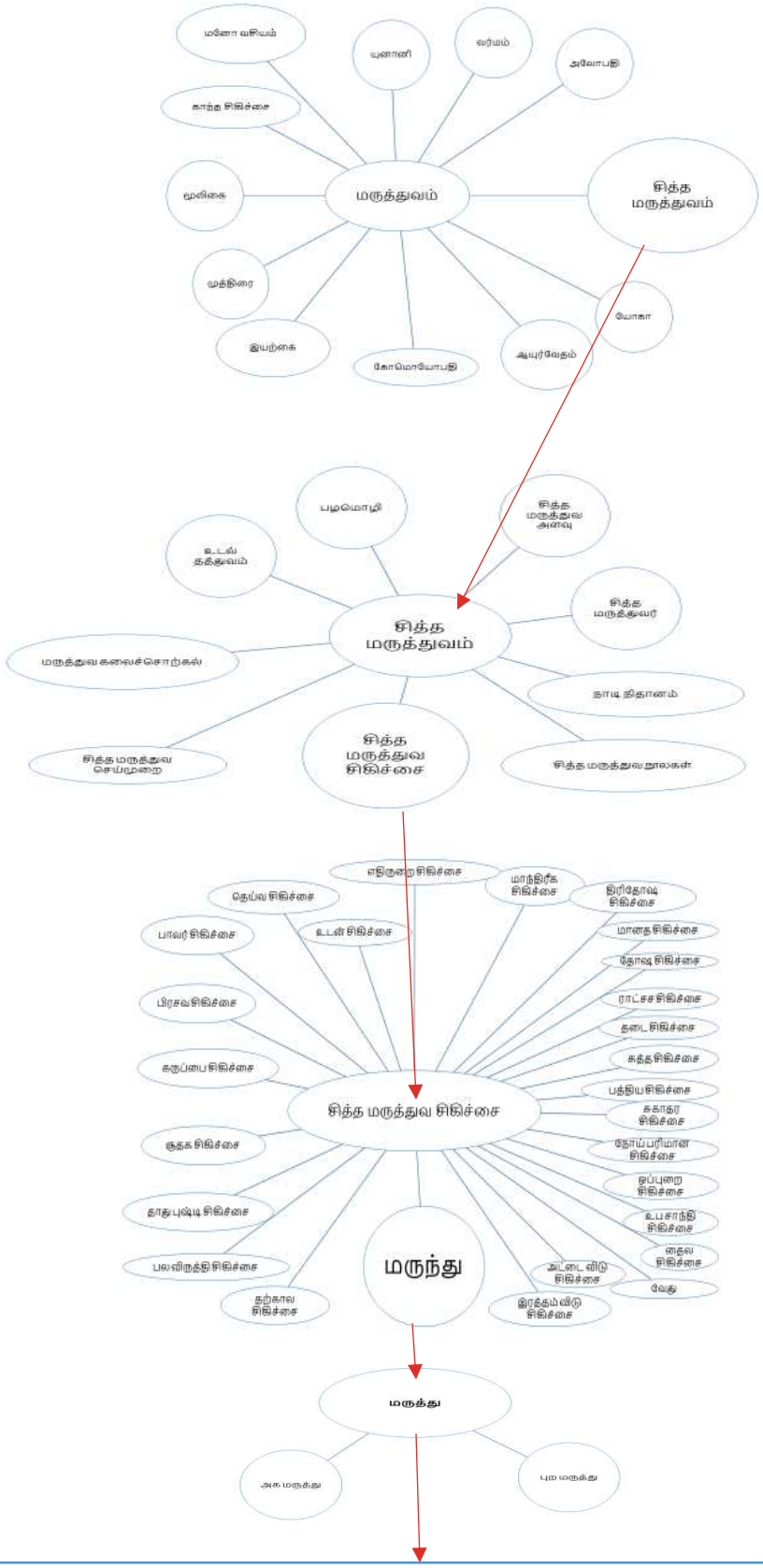
5. Building ontologies from texts using word embedding technology

Old Siddha texts in are available in printed form. Such texts will be digitalized and converted into machine readable format. From such texts we will collect relevant taxonomies or ontologies related to Siddha health care using word embedding technology. In natural language processing (NLP), a word

embedding is a representation of a word. The embedding is used in text analysis. Typically, the representation is a real-valued vector that encodes the meaning of the word in such a way that words that are closer in the vector space are expected to be similar in meaning.

6 User friendly outputs

We propose an HTML based format in which all kinds of lexical relations between health vocabulary will be taken care of.. The example shown below is a sample one. The present structure gives a possibility to expand according to the requirement. The HTML format also allows an ease of access to the data to for information retrieval system. Generating ontology trees from a given huge database structure for a user's query is the backbone activity for the proposed system which is an intelligent information retrieval and visualization system. This type of approach provides a powerful way for data representation and knowledge mining process.



விளக்கம்: சித்த மருத்துவத்தில் தொன்றுதொட்டு உள்ளாட்சி மருந்துகள் 32 வகைகள் எனவும் வெளியாட்சி மருந்து 32 வகைகள் எனவும் கூறப்பட்டுள்ளன. இவையே அகமருந்தென்றும் புறமருந்தென்றும் கூறப்படும்.

The proposed ontological structure based knowledge system for health domain will be of great help to those who are involved in medical practice. The general public can also be benefited by the proposed socio-ontological-cum-knowledge based structure of health domain.

References

Busa, F., Calzolari, N., and Lenci, A. 2001. "Generative Lexicon and SIMPLE Model: Developing Semantic Resources for NLP." In: Bouillon P. and Busa, F. (ed.) The language of Word Meaning, Cambridge University Press, Cambridge, pp. 333-362.

Busa, F., Calzolari, N., Lenci, A. and J. Pustejovsky. 2001. "Building a Semantic Lexicon: Structuring and Generating Concepts." In: H. Bunt et al. (ed.) Computing meaning, Volume 2, 29-51, Kluwer Academic publishers, Netherlands.

iraaceendtiran, ca. 2001. *taRkaalat tamiz coRkaLanjciyam* (a modern Thesaurus for Tamil). Tamil University, Thanjavur.

iraaceendtiran ca. & *paaskaran* ca. 2006 *tamiz mincoRkaLanjciyam* (an electronic Thesaurus for Malayalam). Malayalam University, Thanjavur.

Nida, E.A. 1975a. Componential Analysis of Meaning: An Introduction to Semantic Structure. Mouton, The Hague.

Nida, E.A. 1975.b. Exploring Semantic Structure. The Hague: Mouton.

Rajendran, S. 2001. *taRkaalat tamizc coRkaLanjciyam* [Thesaurus for Modern Malayalam]. Malayalam University, Thanjavur.

Rajendran, S. 1995. "Towards a Compilation of a Thesaurus for Modern Tamil". South Asian Language Review. 5.1:62-99.